

# SUMS OF POSSIBLY ASSOCIATED BERNOULLI VARIABLES: THE CONWAY-MAXWELL-BINOMIAL DISTRIBUTION

JOSEPH B. KADANE

**ABSTRACT.** The study of sums of possibly associated Bernoulli random variables has been hampered by an asymmetry between positive correlation and negative correlation. The Conway-Maxwell Binomial (COMB) distribution and its multivariate extension, the Conway-Maxwell Multinomial (COMM) distribution, gracefully model both positive and negative association. Sufficient statistics and a family of proper conjugate distributions are found. The relationship of this distribution to the exchangeable special case is explored, and two applications are discussed.

## 1. SUMS OF POSSIBLY ASSOCIATED BERNOULLI VARIABLES

There often are reasons to suggest that Bernoulli random variables, while identically distributed, may not be independent. For example, suppose pots are planted with six seeds each, where each pot has seeds from a unique plant, but different pot's seeds came from different plants. Suppose that success of a seedling is well-defined. If genetic similarity is the dominant source of non-independence, it is reasonable to suppose positive association. However, if competition for nutrients and sunlight predominates, association could be negative. Hence, it makes sense to find a functional form that gracefully allows for either positive or negative association.

“Association” here means something more general than correlation. Correlation is a particular measure of association, familiar because of its connection with the normal distribution, and its simple relationship to certain expectations. However, there is no particular reason why correlation should be used in non-normal situations if it has undesirable properties.

The desire for a functional form that allows for both positive and negative association runs into the following familiar fact, which is well-known, but for completeness is proved in Appendix A:

**Proposition 1.** *Suppose  $X_1, \dots, X_m$  have (possibly different) means and variances and common pairwise correlations  $\rho$ . Then  $\rho \geq -1/(m-1)$ .*

There are (at least) three different possible strategies for dealing with the asymmetry between positive and negative correlation revealed by the proposition:

- a) abandon correlation as a measure of association
- b) abandon exchangeability of the Bernoulli random variables

- c) model the sum directly, without fully specifying the distribution of the underlying Bernoulli random variables.

Some light on strategies b) and c) is shed by the following proposition, also proved in Appendix A.

**Proposition 2.** *Let  $P\{S = k\} = p_k \geq 0$ , where  $\sum_{k=0}^m p_k = 1$ . Then there exists a unique distribution on  $X_1, \dots, X_m$  such that  $X_1, \dots, X_m$  are exchangeable, and  $\sum_{i=1}^n X_i$  has the same distribution as does  $S$ .*

Proposition 2 is reassuring with respect to strategy c), since the set of distributions on the  $X$ 's corresponding to an arbitrary distribution on their sum is non-empty. However, it also shows that one can assume exchangeability among the  $X$ 's without restricting the distribution of their sum, so strategy b) is superfluous. (This fact is also a consequence of Galambo's (1978) Theorem 3.2.1.)

The distribution studied in this paper pursues strategies a) and c) simultaneously.

There is a voluminous literature on sums of non-independent Bernoulli random variables. An early paper of Skellam (1948) proposed the beta-binomial distribution, a beta mixture of binomials. Thus, the underlying Bernoulli random variables are exchangeable, so this proposal can model positive association, but not negative association (see also Williams (1975)). Altham (1978) introduces an arithmetic and a multiplicative extension of the binomial distribution, with the intent of modeling non-independence. Both models are exchangeable, and hence are limited in modeling negative association.

Kupper and Haseman (1978) propose an exchangeable model extending the binomial distribution, based on ideas of Bahadur (1961). Once again the model is exchangeable, so the bounds on the common correlation allow a narrow band of negative correlations that rapidly diminish to zero as the sample size increases. Awkwardly, the parameter constraints depend on the data. S.R. Paul (1985, 1987) proposes two models that aim to unify the beta-binomial and the Kupper/Haseman models. Both suffer from the inevitable narrow range of negative correlations possible, and from data-dependent parameter constraints.

Other papers discussing variations on exchangeable extensions of the binomial include Prentice (1986); Madsen (1993); Luceno and DeCeballos (1995); Bowman and George (1995); George and Bowman (1995); George and Kodell (1996); Witt (2004) and Hisakado et al. (2006).

Additionally, there is a paper discussing the sum of not-identically distributed Bernoulli random variables with a common correlation (Gupta and Tao, 2010). They apply their results in the context of multiple testing. Two papers present two-state Markov models, which of course in equilibrium have identically distributed margins, but are not necessarily exchangeable (Viveros et al., 1984; Rudolfer, 1990).

By contrast, Ng (1989) starts with a completely general class of discrete distributions, defined in a complicated sequential scheme. He then specializes it to the exchangeable case, but in general allows for arbitrary dependent structures.

The remainder of this paper is organized as follows: Section 2 introduces the Conway-Maxwell Binomial distribution and displays some of its mathematical properties. Section 3 gives sufficient statistics and discusses a conjugate prior family. Section 4 displays some

examples, and gives expressions for its generating functions. The exchangeable case is examined in Section 5, and some applications are shown in Section 6. Appendix C shows that the results given for the COMB distribute extend to its multivariate generalization, the COMM (Conway-Maxwell-Multivariate) distribution.

## 2. THE CONWAY-MAXWELL BINOMIAL DISTRIBUTION

The binomial distribution, the sum of independent Bernoulli random variables, is extraordinarily useful. Yet there are situations in which the assumption of independence is questionable or unwise. The Conway-Maxwell Binomial distribution (COMB) is a convenient two-parameter family that generalizes the binomial distribution and models both positive and negative association among the Bernoulli summands.

The probability mass function of the COMB distribution is given by

$$(1) \quad P\{W = k\} = \frac{p^k(1-p)^{m-k} \binom{m}{k}^\nu}{S(p, \nu)} \quad k = 0, 1, \dots, m$$

where

$$S(p, \nu) = \sum_{k=0}^m p^k(1-p)^{m-k} \binom{m}{k}^\nu.$$

Here  $0 \leq p \leq 1$  and  $-\infty \leq \nu \leq \infty$  (see Shmueli et al. (2005, eqn. (13))). Of course, when  $\nu = 1$ , the binomial distribution results.

When  $\nu > 1$ , the center of the distribution is upweighted relative to the binomial distribution and the tails downweighted. In the limit as  $\nu \rightarrow \infty$ ,  $W$  piles up at  $m/2$  if  $m$  is even, and at  $\lfloor m/2 \rfloor$  and  $\lceil m/2 \rceil$  if  $m$  is odd. Thus the component Bernoulli random variables are negatively related. Conversely, when  $\nu < 1$ , the tails are upweighted relative to the binomial distribution, and the center downweighted. In the limit as  $\nu \rightarrow -\infty$ , (1) puts all its probability on  $W = 0$  and  $W = m$ , which is the extreme case of positive dependence (all  $X$ 's have the same value). As a consequence, the component Bernoullis are positively related. Thus,  $\nu$  measures the extent of positive or negative association in the component Bernoullis. Figure 1 (from Kadane and Naeshagen (2013)) illustrates these points.

The name ‘‘Conway-Maxwell’’ comes from its relationship to the Conway and Maxwell (1962) generalization of the Poisson distribution, COM-Poisson( $\lambda, \nu$ ):

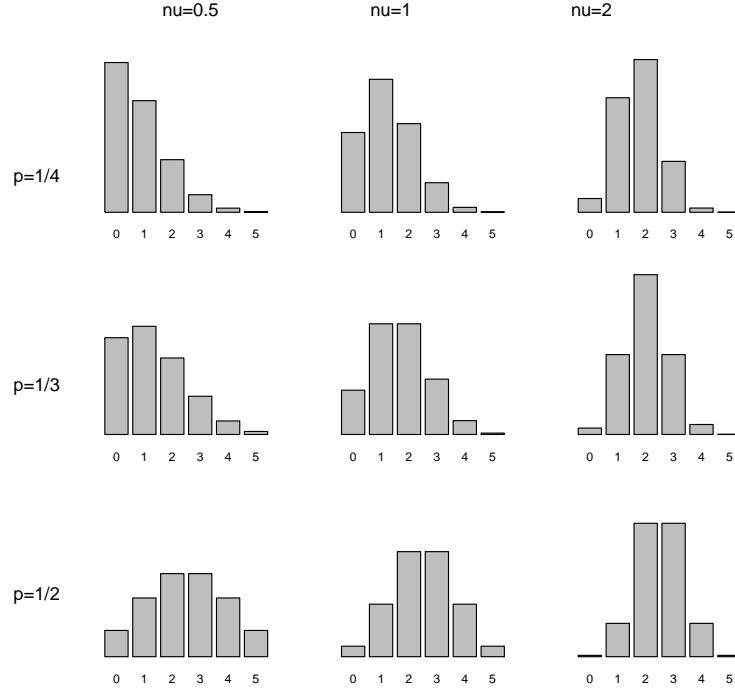
$$(2) \quad P\{W = x\} = \frac{\lambda^x}{(x!)^\nu M(\lambda, \nu)} \quad x = 0, 1, \dots$$

where  $M(\lambda, \nu) = \sum_{j=0}^{\infty} \lambda^j / (j!)^\nu$ .

Shmueli et al. (2005) show that if  $X \sim CMP(\lambda_1, \nu)$  and  $Y \sim CMP(\lambda_2, \nu)$ ,  $X$  and  $Y$  independent, then

$$(3) \quad X \mid X + Y \sim COMB\left(\frac{\lambda_1}{\lambda_1 + \lambda_2}, \nu\right),$$

generalizing the familiar relationship between the Poisson and binomial distributions, when  $\nu = 1$ .

FIGURE 1. Examples of COMB when  $m = 5$ .

### 3. SUFFICIENT STATISTICS AND A CONJUGATE PRIOR FAMILY

Imagine  $n$  samples from a COMB distribution, each with respect to a common  $m$ . Then the likelihood for  $p$  and  $\nu$  is governed by the data  $k_1, \dots, k_n$ , and is given by

$$(4) \quad p(k_1, \dots, k_n \mid p, \nu) = \frac{\prod_{i=1}^n p^{k_i} (1-p)^{m-k_i} \binom{m}{k_i}^\nu}{[S(p, \nu)]^n}$$

The denominator is constant in the data, so it can be ignored. Then

$$\begin{aligned} p(k_1, \dots, k_n \mid p, \nu) &\propto (1-p)^{mn} \prod_{i=0}^m \left( \frac{p}{1-p} \right)^{k_i} \frac{m!^{\nu n}}{(k_i!(m-k_i)!)^\nu} \\ &\propto e^{(\sum_{i=1}^n k_i)(\log(p/(1-p))) - \nu \sum_{i=1}^n \log[k_i!(m-k_i)!]} \\ (5) \quad &= e^{S_1 \log(p/(1-p)) - \nu S_2} \end{aligned}$$

where  $S_1 = \sum_{i=1}^n k_i$  and  $S_2 = \sum_{i=1}^n \log[k_i!(m-k_i)!]$ . Thus the COMB distribution is a member of the exponential family. Consequently, it has a conjugate prior family. To find a convenient form for this family, start over with the likelihood

$$(6) \quad p^k (1-p)^{m-k} \binom{m}{k}^\nu \bigg/ \sum_{k=0}^m p^k (1-p)^{m-k} \binom{m}{k}^\nu.$$

We may take out the inessential factors of  $(1-p)^m(m!)^\nu$ , yielding

$$\left(\frac{p}{1-p}\right)^k \frac{1}{[k!(m-k)!]^\nu} \bigg/ \sum_{k=0}^m \left(\frac{p}{1-p}\right)^k \frac{1}{[k!(m-k)!]^\nu}.$$

Let  $\Psi = \log(p/(1-p))$ . Then the likelihood is

$$(7) \quad \frac{e^{\Psi k}}{[k!(m-k)!]^\nu} \bigg/ \sum_{k=0}^m e^{\Psi k} / [k!(m-k)!]^\nu, \quad k = 0, 1, \dots, m.$$

Consider a conjugate prior of the form

$$(8) \quad \begin{aligned} h(\Psi, \nu) &= g(\Psi, \nu) e^{\Psi a - b\nu} Z^{-c}(\Psi, \nu) K(a, b, c), \\ &-\infty < \Psi < \infty, -\infty < \nu < \infty, \end{aligned}$$

where  $Z(\Psi, \nu) = \sum_{k=0}^m e^{\Psi k} / [k!(m-k)!]^\nu$   
and  $K^{-1}(a, b, c) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(\Psi, \nu) e^{\Psi a - b\nu} Z^{-c}(\Psi, \nu) d\Psi d\nu$ .

To maintain the property that the conjugate prior family is closed under sampling, the factor  $g(\Psi, \nu)$  can be chosen arbitrarily. However, the propriety of the conjugate family is important for two reasons:

- (1) the usual argument for updating prior to posterior in conjugate families depends on the constant of proportionality being finite.
- (2) the proper behavior of numerical algorithms for computing posterior distributions, such as grid methods and Markov Chain Monte Carlo, also depend on the propriety of the posterior.

For those reasons, it makes sense to choose  $g(\Psi, \nu)$  so that  $K^{-1}(a, b, c) < \infty$ . There are additional reasons to do so. The function  $e^{kx}$  goes to infinity as  $x \rightarrow \infty$  if  $k$  is positive, and to infinity as  $x \rightarrow -\infty$  if  $k$  is negative. Hence, if one chose  $g(\Psi, \nu) \equiv 1$  in (8), one would be declaring that extreme values of  $\Psi$  and  $\nu$  are much more likely than others. This would, I judge, be an unusual belief. Instead, I suspect that values of  $\nu$  most likely to be of interest are those close to 1, and values of  $\Psi$  perhaps those close to zero. A simple choice expressing these ideas is

$$(9) \quad g(\Psi, \nu) = \phi(\Psi) \phi(\nu - 1)$$

where  $\phi$  is the normal probability density function. Because the normal distribution goes to zero quickly for values far from its mean, this choice has the implication of “taming” the tails of (8). With this choice, the following theorem results:

**Theorem 1.**  $K^{-1}(a, b, c) < \infty$ .

*The proof is in Appendix B.*

When propriety holds, the updating of (8) with data  $k$  is given by

$$a' = a + k, b' = b + \log(k!(m-k)!), \text{ and } c' = c + 1.$$

## 4. UNDERSTANDING THE COMB DISTRIBUTION

One way to understand a distribution is to look at some representative examples of it. Figure 1 offers a matrix of such examples, for different values of  $p$  and  $\nu$ .

Another way to understand a distribution is by way of its generating functions. These are derived next. Reconsider

$$\begin{aligned}
 S(p, \nu) &= \sum_{k=0}^m \binom{m}{k}^\nu p^k (1-p)^{m-k} \\
 (10) \qquad &= (1-p)^m \sum_{k=0}^m \binom{m}{k}^\nu \left(\frac{p}{1-p}\right)^k \\
 &= (1-p)^m T\left(\frac{p}{1-p}, \nu\right).
 \end{aligned}$$

where  $T(x, \nu) = \sum_{k=0}^m x^k \binom{m}{k}^\nu$ .

Then the probability generating function of the COMB distribution can be expressed as

$$\begin{aligned}
 E(t^x) &= \sum_{k=0}^m t^k p^k (1-p)^{m-k} \binom{m}{k}^\nu / S(p, \nu) \\
 (11) \qquad &= (1-p)^m \sum_{k=0}^m \left(\frac{tp}{1-p}\right)^k \binom{m}{k}^\nu / S(p, \nu) \\
 &= T(tp/(1-p), \nu) / T(p/(1-p), \nu).
 \end{aligned}$$

Similarly, the moment generating function and the characteristic function are, respectively,

$$(12) \qquad E(e^{tx}) = T(e^t p/(1-p), \nu) / T(p/(1-p), \nu)$$

and

$$(13) \qquad E(e^{itx}) = T(e^{it} p/(1-p), \nu) / T(p/(1-p), \nu).$$

## 5. EXCHANGEABILITY

The COMB distribution is a distribution on the sum of  $m$  (possibly dependent) Bernoulli components without specifying anything else about the joint distribution of those components. This section explores the consequences of assuming in addition that those components are exchangeable.

To establish notation, let

$$(14) \qquad p_{i_1, \dots, i_m} = P\{X_1 = i_1, X_2 = i_2, \dots, X_m = i_m\},$$

where each  $i_j \in \{0, 1\}$ . Let  $\pi$  be a permutation of  $(i_1, \dots, i_m)$ . Then the random variables  $X$  are called exchangeable just in case

$$(15) \qquad p_{i_1, \dots, i_m} = p_{\pi(i_1), \pi(i_2), \dots, \pi(i_m)}$$

for all permutations  $\pi$ .

Let  $S(\ell, m)$  be the set of sequences  $(i_1, \dots, i_m)$  with exactly  $\ell$  1's, *i.e.*, satisfying  $\sum_{j=1}^m i_j = \ell$ . There are  $\binom{m}{\ell}$  such sequences in  $S(\ell, m)$ . The following theorem is given in the literature (see Diaconis (1977, Theorem 1) and the references cited there):

**Theorem 2.** *The set  $\mathcal{E}_m$  of exchangeable sequences is a convex set whose extreme points are  $e_0, \dots, e_m$ , where  $e_\ell$  is the measure that puts probability  $1/\binom{m}{\ell}$  on each element of  $S(\ell, m)$  and 0 otherwise. Each point  $x \in \mathcal{E}_m$  has a unique representation as a mixture of the  $m+1$  extreme points.*

Viewed in this light, the exchangeable COMB distribution specifies a particular two parameter family, with parameters  $p$  and  $\nu$ , of weights on the extreme points  $e_0, \dots, e_m$ .

Because  $m$ -exchangeability applies to every permutation of length  $m$ , it implies  $m'$ -exchangeability for each  $m' < m$ . Hence as  $m$  increases,  $m$ -exchangeability becomes increasingly restrictive. In the limit at  $m = \infty$ , deFinetti's Theorem shows that sums of exchangeable random variables are mixtures of Binomial random variables. Because the marginal distribution of each component is Bernoulli, interest centers on the joint distribution of pairs of such variables. By exchangeability, every pair has the same distribution as every other pair, so concentrating on  $(X_1, X_2)$  suffices. Exchangeability implies that  $P\{X_1 = 0, X_2 = 1\} = P\{X_1 = 1, X_2 = 0\}$ , so there are really three probabilities to consider jointly,  $p_{00} = P\{X_1 = 0, X_2 = 0\}$ ,  $p_{01} = p_{10} = P\{X_1 = 0, X_2 = 1\}$ , and  $p_{11} = P\{X_1 = 1, X_2 = 1\}$ . Diaconis (1977, p. 274) introduces a convenient way of graphing these quantities. The graph is reminiscent of barycentric co-ordinates, only here the constraint is slightly different:

$$(16) \quad p_{00} + 2p_{01} + p_{11} = 1 ; p_{ij} \geq 0.$$

Figures 2 and 3 display the possible values of the exchangeable COMB distribution for specified values of  $m$  and  $\nu$ , as  $p$  varies from 0 to 1.

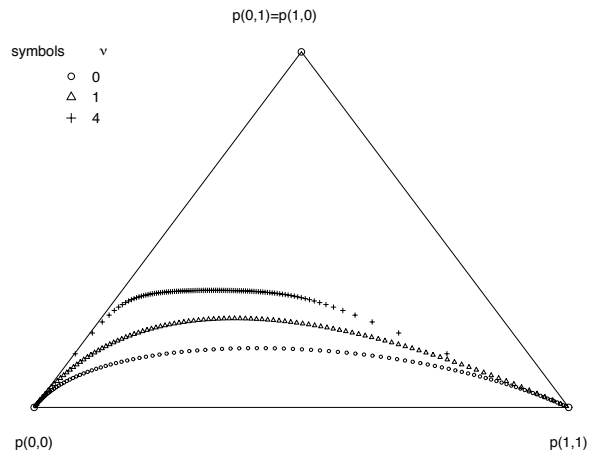
In Figure 2, which is computed at  $m = 3$ , the curve for  $\nu = 4$  is the highest, showing, as expected, more weight on  $p(0, 1) = p(1, 0)$ . The curve  $\nu = 1$  is in the middle; this one corresponds to independence, and is known to be  $y(1 - y)$ . The curve for  $\nu = 0$  is lowest. As  $\nu \rightarrow -\infty$ , this curve descends to the  $p(0, 0)$  to  $p(1, 1)$  lines, indicating that all the probability is at the extremes.

Figure 3 shows the same curve, when  $m = 5$ . The main difference is that the  $\nu = 4$  curve is flatter. Indeed, as  $m \rightarrow \infty$ , this curve will collapse to the  $\nu = 1$  curve.

## 6. APPLICATIONS

### a. An agricultural experiment

Diniz et al. (2010) use a correlated binomial model proposed by Luceno and DeCeballos (1995) to analyze data from an experiment on soybean seeds. The model posits summands that are Binomial  $(m, p)$ , with correlation  $\rho$ . They prove that the sum has the same distribution as a mixture of two distributions: with probability  $(1 - \rho)$  the usual binomial, and with probability  $\rho$ , a Bernoulli  $(p)$  on the points 0 and  $m$ . They use an MCMC with data augmentation to fit the model.

FIGURE 2. Possible values for  $P\{X_1 = i, X_2 = j\}$  when  $m = 3$ .

The data themselves come from having planted six seedlings in each of 20 pots, and using the judgement of an expert as to which seedlings were successful. The goal was to examine the extent to which competition among the seedlings affected the outcomes. The raw data given by Diniz et al. (2010) is reported in Table 1.

# of “good” plants	observed data
0	0
1	2
2	2
3	5
4	5
5	3
6	3

TABLE 1. Observed frequency of “good” plants from Diniz et al. (2010).

To employ the COMB model, I choose to use the prior specified by (9), with  $a = b = c = 0$ . This prior is centered on a Binomial model with  $p = 1/2$  (which implies  $\Psi = 0$ ), which seems reasonable.



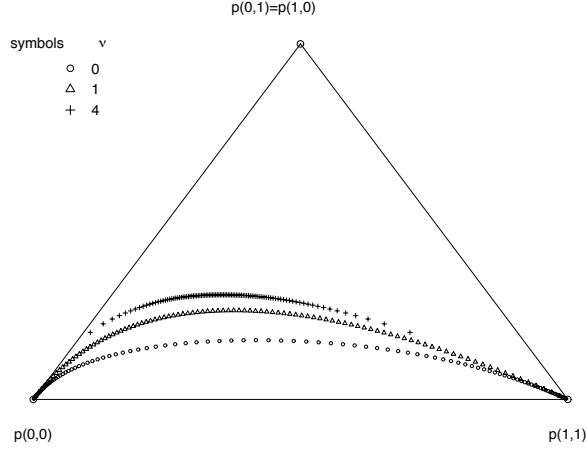


FIGURE 3. Possible values for  $P\{X_1 = i, X_2 = j\}$  when  $m = 5$ .

The contours of the resulting posterior distribution are shown in Figure 4. The maximum posterior point is  $\hat{\Psi} = 0.30$  and  $\hat{\nu} = 0.54$ , with inverse Hessian

$$\Sigma = \begin{pmatrix} 0.028 & 0.018 \\ 0.018 & 0.063 \end{pmatrix}.$$

In view of the elliptical shape of the contours in Figure 4, it is reasonable to approximate the posterior with a normal distribution with mean  $(\hat{\Psi}, \hat{\nu})$  and covariance  $\Sigma$ , as would be suggested by the asymptotic distribution of posterior distributions from conditionally independent models.

Diniz et al. (2010) compare the fit of their model (which they call the “correlated binomial” (CB)), to that of a binomial distribution.

Extending Table 1, Table 2 below reports the estimated fits of all three models:

The sum of squared errors for the three models are as follows: Binomial 8.96; CB 4.16; COMB 3.77.

It is notable that the COMB estimate of  $\nu$  is less than 1, indicating positive association in the soybean seeds. This suggests that competition for nutrients is not the dominant phenomenon in this data set. Further investigation and experimentation might then be warranted to discover the reasons for this positive association. The CB fit did find a

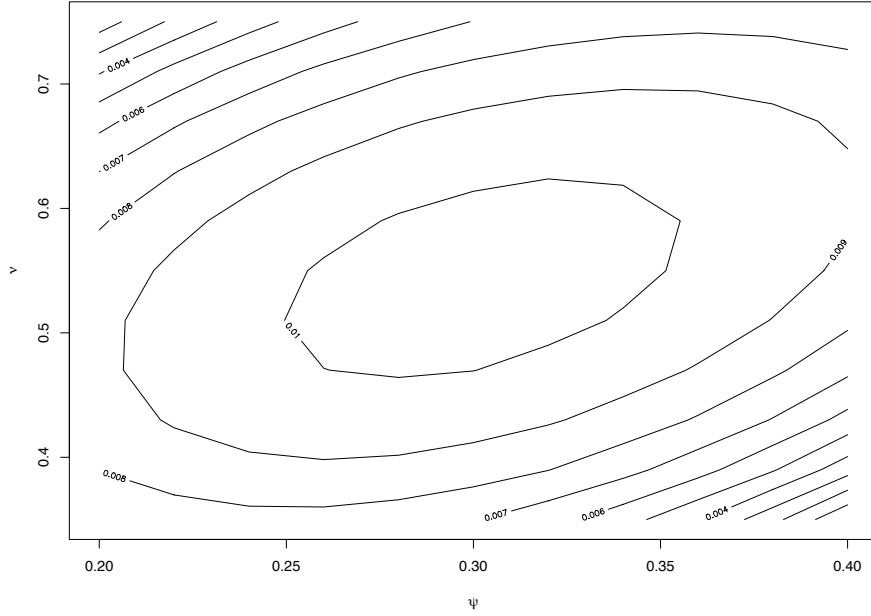


FIGURE 4. Contour plot of the COMB posterior distribution.

# of "good" plants	observed data	binomial fit	CB fit	COMB fit
0	0	0.06	1.19	0.35
1	2	0.61	0.79	1.24
2	2	2.46	2.73	2.76
3	5	5.28	5.03	4.36
4	5	6.37	5.21	5.04
5	3	4.10	2.87	4.14
6	3	1.09	2.17	2.12

TABLE 2. Fits of various models to the soybean data

positive correlation  $\hat{\rho} = 0.13$ . However, the CB model requires  $\rho \geq 0$  and hence it could not have found a negative correlation.

In summary, the COMB model offers the following advantages over the CB model:

- (1) it allows for both positive and negative association
- (2) it allows for a conjugate analysis, obviating the need for an MCMC
- (3) at least for this data set and a squared error metric, it fits better, with the same number of parameters.

### b. Killings in Medieval Norway

In Norway just after the Viking Period, the law distinguished a killing from a murder. In both, there was somebody dead. However, in the former, the killer went to the King's representative within 24 hours and confessed. (Absent such prompt confession, it would be a murder, punishable by execution or banishment). The King's representative would write a letter to the killer stating that the killer was under the protection of the King. An investigation would ensue, resulting in a second letter to the killer, specifying how much was owed to the King, and how much to the family of the deceased. There would then be receipts to the killer for the payments (two more letters), and a final letter from the King's representative to the killer saying that it was all over. Thus the killer would have received five letters.

Several hundred of these letters have survived in the intervening centuries, and a complete list of those found is available. Additionally, there are mentions of killings in other documents such as private letters, Bishop's records, etc. A simple representation of the data is a  $6 \times 2$  matrix, where the first dimension records the number of letters to the killer that survive, and the second is whether or not the killing is mentioned in other sources. Of course, there is the  $(0, 0)$  cell of killings for which no letters survive and for which there are no other mentions. To estimate this cell, and hence the total number of killings, Kadane and Naeshagen (2013, 2014) resort to a dual-systems estimate.

Since there's no obvious reason why the survival of letters in the killer's archive should be related to whether the killing is mentioned in the other sources, an independence assumption between the two dimensions seems reasonable. To model the number of letters from a given killing that might survive, a first thought might be a binomial model. However, since all five letters went to the killer, and were likely stored together, at least at first, it is reasonable to suppose that the event of the survival of a given letter to the killer would be positively associated with the event of the survival of the other letters to the same killer. Thus one would expect  $\nu \leq 1$  in the COMB, and Kadane and Naeshagen imposed a prior on  $\nu$  putting zero probability on the space  $\nu \geq 1$ . As it happened, the data favors  $\nu > 1$ , so the posterior piled up at  $\nu = 1$ , the binomial model.

Nonetheless, this was a successful application of the COMB, in that it allowed for (and rejected) what appeared to be the biggest reasonable threat to the model.

## 7. CONCLUSION

The COMB distribution deserves a place in the tool kit of a statistician. Not all Bernoulli random variables are independent, so a one-parameter extension of the binomial distribution, such as the COM-Binomial, may find other useful applications.

### APPENDIX A. PROOF OF PROPOSITIONS 1 AND 2

Suppose  $X_1, X_2, \dots, X_m$  have the same means and variances, and identical correlations  $\rho$ . Then  $\rho \geq -1/(m-1)$ .

*Proof.* Let  $Y_i = (X_i - E(X_i))/\sigma(X_i)$ ,  $i = 1, \dots, m$ . Then  $Y_1, Y_2, \dots, Y_m$  satisfy  $E(Y_i) = 0$  and  $\text{Var}(Y_i) = 1$ . Because correlations are unaffected by location and scale changes, they

still have common covariance  $\rho$ . Now

$$\begin{aligned}
0 &\leq \text{Var} \left( \sum_{i=1}^m Y_i \right) = E \left( \sum_{i=1}^m Y_i \right)^2 - \left( E \left( \sum_{i=1}^m Y_i \right) \right)^2 \\
&= E \left( \sum_{i=1}^m Y_i \right)^2 = E \sum_{i=1}^m Y_i^2 + \sum_{i=1}^m \sum_{\substack{j=1 \\ i \neq j}}^m E Y_i Y_j \\
&= m + m(m-1)\rho
\end{aligned}$$

from which the desired result follows immediately.  $\square$

*Proof.* For each  $k$ , there are  $\binom{m}{k}$  different arrangements of  $k$   $i$ 's and  $m-k$   $0$ 's. Let each of them have probability  $p_k / \binom{m}{k}$ . Then  $P\{\sum_{i=1}^m X_i = k\} = p_k$  and the  $X$ 's are exchangeable.

To show uniqueness, if the sum of the probabilities of the sequences with exactly  $k$   $1$ 's is not  $p_k$ , the sum condition is violated. If their probabilities are not equal, exchangeability is violated.  $\square$

## APPENDIX B. PROOF OF THEOREM

To obtain an upper bound on  $K^{-1}$ , a lower bound on  $Z$  is needed. I proceed from Jensen's inequality, using the convexity of  $\log(x)$  (*i.e.*, the second derivative is negative):

Let  $q_0, \dots, q_m$  be arbitrary probabilities that are non-negative and sum to 1.

Then

$$\log \left( \sum_{k=0}^m q_k a_k \right) \geq \sum_{k=0}^m q_k \log a_k.$$

Now

$$\begin{aligned}
\log Z(\Psi, \nu) &= \log \sum_{k=0}^m q_k e^{\Psi k} / q_k (k!(m-k)!)^\nu \\
&\geq \sum_{k=0}^m q_k \log \left( e^{\Psi k} / q_k (k!(m-k)!)^\nu \right) \\
&= \Psi \sum_{k=0}^m q_k \cdot k - \nu \sum_{k=0}^m q_k \log(k!(m-k)!) - \sum_{k=0}^m q_k \log q_k.
\end{aligned}$$

Let  $Q$  be a random variable on the non-negative integers  $\{0, 1, \dots, m\}$  with probability mass  $Pr\{Q = k\} = q_k$ . Then the bound can be written as

$$Z(\Psi, \nu) \geq \underline{Z}(\Psi, \nu) = e^{\Psi E(Q) - \nu E(\log(Q!(m-Q)!))} \prod_{k=0}^m q_k^{q_k}$$

Therefore

$$\begin{aligned} K^{-1}(a, b, c) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(\Psi, \nu) e^{a\Psi - b\nu} Z^{-c}(\Psi, \nu) d\Psi d\nu \\ &\geq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(\Psi, \nu) e^{a\Psi - b\nu} e^{c\{\Psi EQ - \nu E(\log(Q!(m-Q)!))\}} \prod_{k=0}^m q_k^{cq_k}. \end{aligned}$$

Substituting  $g(\Psi, \nu) = \phi(\Psi)\phi(\nu - 1)$  and collecting terms

$$K^{-1}(a, b, c) \geq \int_{-\infty}^{\infty} \frac{e^{-\Psi^2/2 + \Psi(a - cE(Q)) - (\nu-1)^2/2 - \nu(b - cE(\log Q!(m-\theta)!))}}{2\pi} \prod_{k=0}^m q_k^{cq_k} d\Psi d\nu.$$

Both the integral with respect to  $\Psi$  and that with respect to  $\nu$  are of the form  $e^{-(x^2/2) + xk}$ , which are normal integrals, and hence finite. The constant  $(\prod_{k=0}^m q_k^{cq_k})/2\pi$  is also finite. Therefore we have  $K^{-1}(a, b, c) < \infty$ , as was to be shown.  $\square$

Remarks:

- (1) This proof uses bounds similar to those in Kadane et al. (2006).
- (2) This theorem also holds if instead of  $g(\theta, \nu)$  as specified in (9), any other normal distribution for  $(\Psi, \nu)$  were used instead.

#### APPENDIX C. THE CONWAY-MAXWELL-MULTIVARIATE DISTRIBUTION

The Conway-Maxwell-Multivariate (COMM) Distribution has probability mass function (for fixed  $m$ )

$$P\{\mathbf{X} = \mathbf{k} | (\mathbf{P}, V)\} = \binom{m}{\mathbf{k}}^\nu \prod_{i=1}^r p_i^{k_i} / \sum_{\mathbf{j} \in D} \binom{m}{\mathbf{j}}^\nu \prod_{i=1}^r p_i^{j_i}, \mathbf{k} \in D$$

where

$$\begin{aligned} \mathbf{p} &= (p_1, \dots, p_r), p_i \geq 0, \sum_{i=1}^r p_i = 1 \\ \mathbf{k} &= (k_1, \dots, k_r), k_i \geq 0, \sum_{i=1}^r k_i = m, k_i' \text{ s integers} \\ \binom{m}{\mathbf{k}} &= \frac{m!}{k_1! k_2! \dots k_r!} \end{aligned}$$

and  $D$  is the set of vectors of integers  $\mathbf{j}$  satisfying  $j_i \geq 0$  and  $\sum_{i=1}^r j_i = m$ .

**Proposition 3.** Suppose  $X_1, \dots, X_r$  are independently distributed with probability mass function Conway-Maxwell Poisson  $X_i \sim CMP(\lambda_i, \nu)$ :

$$P\{X_i = s_i | \lambda_i, \nu\} = \frac{\lambda_i^{s_i}}{(s_i!)^\nu Z(\lambda_i, \nu)}$$

where  $Z(\lambda_i, \nu) = \sum_{j=0}^{\infty} \frac{\lambda_i^j}{(j!)^\nu}$ .

Then  $\mathbf{X} | \sum_{i=1}^r X_i = m$  has a COMM Distribution with parameters  $p_i = \lambda_i/\lambda$  and  $\nu$ , where  $\lambda = \sum_{i=1}^r \lambda_i$ .

*Proof.* Let  $S = \sum_{i=1}^r X_i$ ,  $\lambda = (\lambda_1, \dots, \lambda_r)$  and  $G(\mathbf{p}, \nu) = \sum_{\mathbf{j} \in D} \binom{m}{\mathbf{j}}^\nu \prod_{i=1}^r p_i^{j_i}$ .

Then

$$\begin{aligned} P\{S = m\} &= \sum_{\mathbf{j} \in D} \prod_{i=1}^r \frac{\lambda_i^{j_i}}{(j_i!)^\nu Z(\lambda_i, \nu)} \\ &= \frac{\lambda^m}{(m!)^\nu} \prod_{i=1}^r Z(\lambda_i, \nu) \sum_{\mathbf{j} \in D} (\lambda_i/\lambda)^{j_i} \binom{m}{\mathbf{j}}^\nu \\ &= \frac{\lambda^m}{(m!)^\nu} \cdot \frac{1}{\prod_{i=1}^r Z(\lambda_i, \nu)} G(\mathbf{x}/\lambda, \nu) \end{aligned}$$

Hence

$$\begin{aligned} P\{\mathbf{X} = \mathbf{k} \mid S = m\} &= \prod_{i=1}^r \frac{\lambda_i^{k_i}}{(k_i!)^\nu} Z(\lambda_i, \nu) \bigg/ \frac{\lambda^m}{(m!)^\nu} \prod_{i=1}^r Z(\lambda_i, \nu) \\ &= \prod_{i=1}^r (\lambda_i/\lambda)^{k_i} \binom{m}{\mathbf{k}}^\nu \bigg/ G(\lambda/\lambda, \nu), \text{ for } \mathbf{k} \in D \end{aligned}$$

which is the probability mass function of the COMM distribution with parameters  $p_i = \lambda_i/\lambda$  ( $i = 1, \dots, r$ ) and  $\nu$ .  $\square$

**Proposition 4.** Let  $P\{\mathbf{S} = \mathbf{k}\} = p_{\mathbf{k}} \geq 0$ , where  $\sum_{\mathbf{k} \in D} p_{\mathbf{k}} = 1$ . Then there exists a unique distribution on  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m$  such that  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m$  are exchangeable and  $\sum_{i=1}^m \mathbf{X}_i$  has the same distribution as does  $S$ .

*Proof.* For each  $\mathbf{k} \in D$ , there are  $\binom{m}{\mathbf{k}}$  different arrangements of 1's and 0's such that each vector component  $i$  has  $k_i$  1's and  $(m - k_i)$  0's. Let each such arrangement have probability  $p_{\mathbf{k}}/\binom{m}{\mathbf{k}}$ . Then  $P(\sum_{i=1}^r \mathbf{X}_i = \mathbf{k}) = p_{\mathbf{k}}$  and the  $\mathbf{X}_i$ 's are exchangeable. To show uniqueness, if the sum of the probabilities of the sequences with  $k_i$  1's in the  $i^{\text{th}}$  vector component for each  $i$  were not  $p_{\mathbf{k}}$ , the sum constraint would not be met. If they did not have equal probability, exchangeability would be violated.  $\square$

**Proposition 5.** The COMM distribution has the following sufficient statistics:

$$\begin{aligned} S_0 &= \sum_{j=1}^n \log[k_{1j}! \dots k_{rj}!] \\ S_i &= \sum_{j=1}^n k_{ij}, i = 1, \dots, r-1 \end{aligned}$$

where  $k_{ij}$  is the  $i^{\text{th}}$  component of the  $j^{\text{th}}$  sample.

*Proof.*

$$\begin{aligned} p(\mathbf{k}_1, \dots, \mathbf{k}_n \mid \mathbf{p}, \nu) &= \prod_{j=1}^n \left[ \binom{m}{\mathbf{k}_j}^\nu \prod_{i=1}^r p_i^{k_{ij}} \bigg/ G(\mathbf{p}) \right] \\ &\propto p_r^{nm} (m!)^{\nu n} \prod_{j=1}^n \left[ \prod_{i=1}^{r-1} (p_i/p_r)^{k_{ij}} \right] [\prod_{i=1}^r k_{ij}!]^{-\nu} \\ &\propto e^{\sum_{i=1}^{r-1} \log(p_i/p_r) \sum_{j=1}^n k_{ij} - \nu \sum_{j=1}^n \log(\prod_{i=1}^r k_{ij}!)} \\ &= e^{\sum_{i=1}^{r-1} \log(p_i/p_r) S_i - \nu S_0}. \end{aligned}$$

$\square$

Let  $\boldsymbol{\psi} = (\log(p_1/p_r), \log(p_2/p_r), \dots, \log(p_{r-1}/p_r))$ . Consider a conjugate family of the form

$$h^*(\boldsymbol{\psi}, \nu) = g(\boldsymbol{\psi}, \nu) e^{-\boldsymbol{\psi} \cdot \mathbf{a} - b\nu} G^{-c}(\boldsymbol{\psi}, \nu) K(\mathbf{a}, b, c)$$

where  $\mathbf{a}$  is a vector of length  $r - 1$ , and  $b$  and  $c$  are positive numbers. Here  $G(\boldsymbol{\psi}, \nu) = \sum_{\mathbf{j} \in D} \binom{m}{\mathbf{j}}^\nu \prod_{i=1}^{r-1} \psi_i^{j_i}$ . If  $g(\boldsymbol{\psi}, \nu)$  is taken to have a normal distribution (of dimension  $r$ ), then  $K^{-1}(\mathbf{a}, b, c) < \infty$ .

In this case, updating occurs as follows:

$$\mathbf{a}' = \mathbf{a} + \mathbf{k}^*, b' = b + \log(k_1! k_2! \dots k_r!), c' = c + 1,$$

where  $\mathbf{k}^* = (k_1, k_2, \dots, k_{r-1})$ .

*Proof.* (Generalization of Appendix B).

Let  $Q$  be a distribution over  $D$  and let  $Q^* = (Q_1, \dots, Q_{r-1})$ . Then Jensen's inequality gives

$$\begin{aligned} \log \sum_{\mathbf{k} \in D} q_{\mathbf{k}} a_{\mathbf{j}} &\geq \sum_{\mathbf{k} \in D} q_{\mathbf{k}} \log a_{\mathbf{k}} \\ \log G(\boldsymbol{\psi}, \nu) &= \log \sum_{\mathbf{j} \in D} q_{\mathbf{j}} \log(e^{\boldsymbol{\psi} \cdot \mathbf{k}^*} / q_{\mathbf{k}} (\prod k_i!)^\nu) \\ &\geq \sum_{\mathbf{j} \in D} q_{\mathbf{j}} \log(e^{\boldsymbol{\psi} \cdot \mathbf{k}^*} / q_{\mathbf{k}} (\prod k_i!)^\nu) \\ &= \boldsymbol{\psi} \cdot \sum_{\mathbf{j} \in D} q_{\mathbf{j}} \mathbf{k}^* - \nu \sum_{\mathbf{k} \in D} q_{\mathbf{k}} \log(\prod k_i!) - \sum_{\mathbf{j} \in D} q_{\mathbf{j}} \log q_{\mathbf{j}} \\ &= \boldsymbol{\psi} E(Q^*) - \nu E \log(Q_1! Q_2! \dots Q_r!) - \sum_{\mathbf{j} \in D} q_{\mathbf{j}} \log q_{\mathbf{j}}. \end{aligned}$$

Since these are linear in  $\boldsymbol{\psi}$  and  $\nu$ , with any normal prior on  $(\boldsymbol{\psi}, \nu)$ , the integral  $K^{-1}$  is finite  $\square$

#### ACKNOWLEDGEMENTS

I thank Christian Robert, Kim Sellers, Galit Shmueli and Rebecca Steorts for helpful comments.

#### REFERENCES

- Altham, P. (1978). “Two generalizaitons of the binoial distribution.” *J. Roy. Stat. Soc. C*, 27, 2, 162–167.
- Bahadur, R. (1961). “A representation of the joint distribution of  $n$  dichotomous items.” In *Studies in Item Analysis and Predictions*, ed. H. Solomon. Stanford, California: Stanford University Press.
- Bowman, D. and George, E. (1995). “A saturated model for analyzing exchangeable binary data: Applications to clinical and developmental toxicity studies.” *JASA*, 90, 871–879.
- Conway, R. and Maxwell, W. (1962). “A queuing model with state dependent service rates.” *Journal of Industrial Engineering*, 12, 132–136.
- Diaconis, P. (1977). “Finite forms of de Finetti's Theorem on exchangeability.” *Synthese*, 36, 271–281.
- Diniz, A., Tutia, M., and Laite, J. (2010). “Bayesian analysis of a correlated binomial model.” *Brazilian Journal of Probability and Statistics*, 24, 1, 68–77.
- Galambos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics*. New York: J. Wiley and Sons.

- George, E. and Bowman, D. (1995). "A full likelihood procedure for analysing exchangeable binary data." *Biometrics*, 51, 512–523.
- George, E. and Kodell, R. (1996). "Tests of independence, treatment heterogeneity and dose-related trend with exchangeable binary data." *JASA*, 91, 1602–1610.
- Gupta, R. and Tao, H. (2010). "A generalized correlated binomial distribution with application in multiple testing." *Metrika*, 71, 59–77.
- Hisakado, M., Kitsukawa, K., and Mori, S. (2006). "Correlated binomial models and correlation structures." *Journal of Physics A: Mathematical and General*, 39, 15365.
- Kadane, J. and Naeshagen, F. (2013). "The number of killings in southern rural Norway, 1300–1569." *Annals of Applied Statistics*, 7, 2, 846–859.
- (2014). "The rate of killings in southern rural Norway, 1300–1569." *Scandinavian Journal of History*, To appear.
- Kadane, J., Shmueli, G., Minka, T., Borle, S., and Boatwright, P. (2006). "Conjugate analysis of the Conway-Maxwell-Poisson Distribution." *Bayesian Analysis*, 1, 363–374.
- Kupper, L. and Haseman, J. (1978). "The use of a correlated binomial model for the analysis of certain toxicological experiments." *Biometrics*, 34, 69–76.
- Luceno, A. and DeCeballos, F. (1995). "Describing extra-binomial variation with partially correlated models." *Communications in Statistics: Theory and Methods*, 24, 6, 1637–1653.
- Madsen, R. (1993). "Generalized binomial distributions." *Communications in Statistics: Theory and Methods*, 22, 11, 3065–3086.
- Ng, T.-H. (1989). "A new class of modified binomial distributions with applications to certain toxicological experiments." *Communications in Statistics: Theory and Methods*, 18, 9, 3477–3492.
- Paul, S. (1985). "A three-parameter-generalization of the binomial distribution." *Communications in Statistics, Theory and Methods*, 14, 6, 1497–1506.
- (1987). "On the beta-correlated binomial distribution - A three parameter generalization of the binomial distribution." *Communications in Statistics, Theory and Methods*, 16, 5, 1473–1478.
- Prentice, R. (1986). "Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors." *JASA*, 81, 321–327.
- Rudolfer, S. (1990). "A Markov chain model of extrabinomial variation." *Biometrika*, 77, 2, 255–264.
- Shmueli, G., Minka, T., Kadane, J., Borle, S., and Boatwright, P. (2005). "A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution." *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 54, 1, 127–142.
- Skellam, J. (1948). "A probability distribution derived from the binomial distribution by regarding the probability of success as variable between sets of trials." *J. Royal Statist. Soc. B*, 10, 257–261.
- Viveros, R., Balasubramanian, K., and Balakrishnan, N. (1984). "Binomial and negative binomial analogues under correlated Bernoulli trials." *The American Statistician*, 48, 3, 243–247.



- Williams, D. (1975). “The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity.” *Biometrics*, 31, 949–952.
- Witt, G. (2004). *Moody’s correlated binomial default distribution*. Moody’s Investor Service.